LINGUISTIC ANALYSIS

EXAMPLE 1:

"text": "when is the next train from Winterstrae-12 to kieferngarten",

"intent": "FindConnection"

when ADV

is AUX

the next train NOUN

from ADP

Winterstrae-12 PROPN

to ADP

Kieferngarten? PROPN

EXAMPLE 2:

"text": "when does the next u-bahn leaves from Garching-forschungszentrum?",

"intent": "DepartureTime"

when ADV

does AUX

the next U-bahn leaves NOUN

from ADP

Garching-forschungszentrum? PROPN

Analysis:

Example 1 and 2 have different Intents.

But when we are analysing the linguistic features, the syntactic structure of both are similar.

1 : ADV-AUX-NOUN-ADP-PROPN-ADP-PR0PN

2 : ADV-AUX-NOUN-ADP-PROPN

Also, the starting ADJ is also same. 'When<ADJ>' denotes the temporal aspect, so can easily be confused as DepartureTime.

This feature is most important consideration while doing error-analysis for bigger datasets

So the syntactic feature at last part of sentence ADP<to>-PROPN<GPE-entity> is most significant bit.

Cross-Lingual / Multi-lingual dependency

Language-Dependent Modules :

Preprocessing Module:

1. Text Cleaning
2. Tokenization
3. Word2Vec Embeddings

Language-Independent Modules :

1. Building model
2. Training
3. Prediction

Code-Mixed Data

Issues:

1. Romanized Dataset
2. Contractions

   Ex: ”between”→”btwn” ; ”bhut bdiya”
3. Non-grammatical constructions

   Ex: ”sirji hlp plzz naa”

Solution (New & Open-ended Research problem):

1. Language Identification Module

2. Cross-Lingual word embeddings & Parallel corpus

3. Micro-word architecture for finding language-mix in a single word

   Ex: Sirji Sir: en and ji: Hin

4. Hybrid approaches using DL + Linguistics

Sparse - Data Problem

More training data can be generated using:

1. Data Augmentation (using similarity metrics)
2. Template-based data generation
3. Synthetic data using SMOTE
4. Transfer Learning Techniques